ThinkFlow: Self-Evolving Probabilistic Latent Memory for Lifelong Conversational Agents

Cai Ke[1,2], Xin Liu[1*], Han Zhang[1], Jiangyue Yan[1,2], Zike Yuan[1,2], Ling Deng[3], Yue Yu[1], Hui Wang[1], and Ruifeng Xu[2,1*]
[1]Pengcheng Laboratory, China [2]Harbin Institute of Technology, Shenzhen, China
[3]China Unicom Greater Bay Area Innovation Institute, China
kecai@stu.hit.edu.cn, xuruifeng@hit.edu.cn

Abstract

Lifelong conversational agents rely on memory systems to maintain deep, context-aware interactions with users. However, existing explicit textual memory pipelines suffer from a severe information bottleneck, often losing subtle behavioral patterns and emotional shifts. Furthermore, being typically static post-deployment, they cannot autonomously adapt to personal habits and preferences without manual feedback. Cognitive science, however, suggests that humans maintain mental models purely in a latent space and continuously refine them through predictive coding. Inspired by this, we propose **ThinkFlow**, a novel end-to-end latent memory framework for lifelong conversational agents. ThinkFlow bypasses the text bottleneck by dynamically compressing conversational flows into probabilistic latent memory skills, autonomously consolidating complex user states into disentangled, continuous vectors without semantic interference. To break this barrier, we introduce a test-time evolution paradigm. By coupling teacher-guided latent alignment to bootstrap the initial state with a self-supervised next-user-utterance prediction task for continuous refinement, the framework successfully overcomes cold-start challenges and achieves label-free lifelong personalization. Extensive experiments on long-term conversation benchmarks demonstrate that ThinkFlow significantly outperforms prevailing memory systems, providing highly personalized and contextually accurate responses over extended multi-session interactions.

1 Introduction

Recent advances in large language models (LLMs) have transformed them from static question-answering tools into conversational agents expected to build engaging, lifelong relationships with users (Xu et al., 2022; Bae et al., 2022; Zhang et al., 2023; Lu et al., 2023; Jang et al., 2024; Zhong et al., 2024; Li et al., 2025a; Liang et al., 2026; Ke et al., 2026a). This evolution has catalyzed a paradigm shift toward personalized memory retrieval (Ong et al., 2025; Tan et al., 2025; Zhang et al., 2025; Jiang et al., 2025a,b; Zhang et al., 2026b), which focuses on capturing user-specific histories, evolving preferences, and highly contextualized interactions to tailor responses to a particular user at a specific moment. However, in real-world scenarios, standard turn-by-turn interactions inevitably suffer from episodic amnesia, failing to satisfy the demand for deep, persistent personalization over prolonged multi-session engagements (Maharana et al., 2024; Wu et al., 2025). Simply feeding the entire raw multi-session dialog history directly into the context window is not only computationally prohibitive but also severely exacerbates the semantic dilution problem, leading to retrieval degradation (Liu et al., 2024a; Xu et al., 2025a; Laban et al., 2026). This underscores the critical need for an efficient, dynamic memory mechanism capable of managing lifelong human-computer companionship while maintaining scalable personalization.

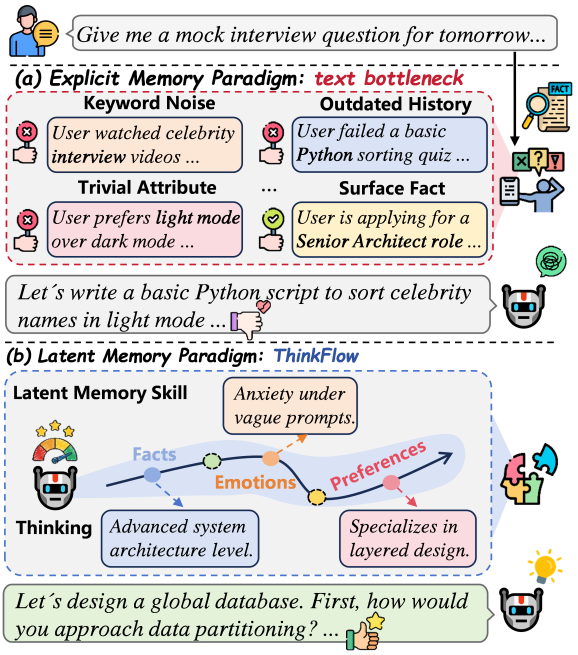

Figure 1: Conceptual overview of our **ThinkFlow**, mitigating the text bottleneck (e.g., Keyword Noise) of explicit memory via dynamic latent memory skills.

The current paradigm primarily relies on explicit text-

* Corresponding authors.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 14126–14141
October 24-29, 2026 ©2026 Association for Computational Linguistics

based memory systems, utilizing complex pipelines to manage the entire memory lifecycle from the construction and management to the retrieval of historical conversations. While yielding remarkable results, these explicit methods fundamentally suffer from a severe information bottleneck (Packer et al., 2023; Liu et al., 2024b; Mei et al., 2024; Wang et al., 2025a; Chhikara et al., 2025; Xu et al., 2025b; Kang et al., 2025; Fang et al., 2026). By forcing the rich, continuous flow of human interaction to be materialized into rigid, discrete plain text, they inherently strip away subtle nuances including implicit user preferences, shifting emotional states, and unspoken behavioral patterns. Furthermore, these text-based memory banks are typically static post-deployment and lack the ability to autonomously adapt to user idiosyncrasies without explicit feedback. This limitation prevents LLMs from developing a deep, seamlessly evolving understanding of the user, keeping them from providing truly empathetic and human-like lifelong companionship. This points towards the need for a memory mechanism that, much like human cognition, operates implicitly and continuously self-evolves without relying on explicit text generation.

According to Cognitive Science, specifically the *Implicit Theory of Mind* (Apperly and Butterfill, 2009; Schneider et al., 2012), humans do not memorize discrete text summaries of past interactions; rather, they naturally and unconsciously track others' mental states, continuously updating their internal mental models purely in a latent, abstract space. Furthermore, based on *Predictive Coding* (Clark, 2013; Millidge et al., 2021), the human brain constantly generates top-down predictions of future sensory inputs and refines its internal models based on prediction errors. As illustrated in Figure 1 (a), traditional explicit memory systems suffer from the text bottleneck, where complex behavioral nuances are lost during summarization, leading to rigid and static profiles. In contrast, as shown in Figure 1 (b), our proposed latent memory dynamically compresses the continuous conversational flow into disentangled skill vectors (e.g., Facts, Emotions, Preferences). Driven by implicit feedback from predicting the user's next action, these latent skills autonomously update and evolve. **Therefore, we argue that by abandoning the explicit text bottleneck and maintaining a continuously evolving memory purely in the latent space through self-supervised predictive feedback, we can achieve true, zero-shot personalization in long-term conversations.**

To reach this goal, we propose **ThinkFlow**, an end-to-end latent memory framework designed to autonomously extract, filter, and align historical information. Specifically, to overcome the text bottleneck, we introduce *Probabilistic Latent Memory Skills (PLMS)*, which compress redundant context into K dense, disentangled skill vectors that act as specialized cognitive receptors. To mitigate semantic dilution and filter out conversational noise, a *Gated Latent Consolidator (GLC)* strictly controls the updating of these skills. Subsequently, a *Context-Aware Hyper-Aligner (CAHA)* dynamically bridges the distributional gap between historical latent skills and the current semantic context of the LLM. Most importantly, to break the deployment bottleneck of static models, we introduce a *Self-Supervised Test-Time Evolution* paradigm. After an initial Teacher-Guided Latent Alignment, ThinkFlow engages in continuous lifelong learning via a *Next-User-Utterance Prediction* task. It utilizes the user's implicit reaction as an abundant, free supervisory signal to refine its latent skills based on prediction errors. Experimental results on long-term conversation benchmarks demonstrate that ThinkFlow significantly enhances personalized response generation and memory utilization. **The contributions of this work can be summarized as follows:**

- We explore a new end-to-end latent memory paradigm for lifelong conversational agents, bypassing the traditional explicit text bottleneck to effectively prevent information loss during long-term interactions.
- We are the first to dynamically compress the continuous conversational flow into disentangled skill vectors, enabling the agent to autonomously filter, extract, and consolidate complex user states without semantic interference.
- We introduce a novel self-supervised test-time evolution paradigm. By leveraging implicit predictive feedback during interactions, the proposed framework effectively overcomes the initial cold-start hurdle and achieves continuous, label-free lifelong personalization.
- Experimental results demonstrate the superiority of ThinkFlow over prevailing memory systems, showcasing its ability to provide highly personalized and contextually accurate responses in extended multi-session interactions.

2 Related Work

Explicit Memory Paradigm. Long-term memory is critical for maintaining personalization and coherence in lifelong conversational agents. Existing systems predominantly rely on explicit textual memory, which often starts with *structured retrieval-augmented generation* to organize past conversations into structured topologies like hierarchical trees or knowledge graphs for context-aware reasoning (Sarthi et al., 2024; Edge et al., 2024; Rezazadeh et al., 2025). Moving beyond static structures, *memory-augmented generation* paradigms dynamically extract discrete facts or generate condensed textual summaries of previous sessions to build plug-and-play memory banks (Lu et al., 2023; Zhong et al., 2024; Chen et al., 2025; Li et al., 2025a; Wang et al., 2025b; Ke et al., 2025; Ong et al., 2025; Fang et al., 2026). To actively govern these records, modern frameworks transition to *agentic memory management*, employing

autonomous agents or hierarchical storage architectures to read, write, and update these explicit textual blocks (Packer et al., 2023; Xu et al., 2025b; Chhikara et al., 2025; Kang et al., 2025; Ke et al., 2026a,b). **Different from these methods that rely on explicit textual pipelines and suffer from severe information loss or redundancy, our ThinkFlow bypasses this text bottleneck entirely by compressing and evolving conversational flows purely in the latent space as continuous, probabilistic memory skills**.

Latent Memory Paradigm. To mitigate the token costs and information bottlenecks of explicit textual memory, latent space has emerged as a promising memory substrate due to its token efficiency, machine-native representation, and end-to-end learnability (Hu et al., 2025; Yu et al., 2026). For textual agents, pioneer works encode multi-session experiences into continuous vector spaces to support agentic reasoning, such as representing historical sessions as continuous trajectory embeddings (Zhang et al., 2026c) or caching key contextual patterns into latent memories (Xu et al., 2025c; Li et al., 2025b; Liu et al., 2025; Zhang et al., 2026a). **Different from existing latent memory methods that rely on static latent vectors, ThinkFlow models memory as probabilistic memory skills to capture interaction uncertainty. Furthermore, our method continuously predicts the user's next response to capture their implicit feedback and shifting preferences, achieving self-evolution during online conversations and enabling label-free lifelong personalization**.

3 Methodology

Drawing inspiration from the *Implicit Theory of Mind* in cognitive science, humans naturally and unconsciously track others' mental states, selectively consolidating salient interaction histories to continuously update their internal mental models purely in the latent space. Inspired by this, to address the information loss caused by the traditional text-based memory bottleneck, we propose ThinkFlow. Instead of saving chat history as discrete text, ThinkFlow operates purely in the latent space by compressing the continuous thinking flow into probabilistic memory skills. For a conversation at turn t, let x_t be the user query, a_t be the agent response, $H_{<t}$ be the text history, and $M_t \in \mathbb{R}^{K\times D}$ be the latent memory state comprising K skill slots and D dimensions.

3.1 Overall Architecture

At the core of this architecture is the concept of *latent memory skills*. Instead of a monolithic hidden state, we decompose the memory into multiple continuous skill vectors. This skill-centric design is crucial, as it allows the system to autonomously capture and decouple diverse aspects of the user's profile—such as factual events, emotional states, and implicit preferences—into specialized channels. The architecture comprises three interconnected components.

Probabilistic Latent Memory Skills. To extract fine-grained semantics from the redundant conversational context, the PLMS compresses the thought stream (hidden states) $E_t \in \mathbb{R}^{L\times D}$ of the current turn into K dense memory skills. These skills act as specialized cognitive receptors. To capture the intricate dependencies between the predefined memory slots and the input sequence, we apply cross-attention, where a set of learnable latent memory skills $S \in \mathbb{R}^{K\times D}$ attends to E_t. This skill-based design is crucial for long-term agents: it enforces a disentangled memory skill representation where different skills autonomously specialize in distinct facets of the user profile, such as factual events, personal preferences, or emotional states. By doing so, it prevents semantic interference among diverse conversational topics and allows the agent to precisely retrieve specific memory skill fragments when needed.

To prevent the variance of the compressed skills from collapsing into a deterministic point mass, we introduce a distribution-preserving scaling factor. The scaled attention output is defined as:

$$C_t = \Big(\text{Softmax}\big(\frac{SE_t^T}{\sqrt{D}}\big)E_t\Big) \times \frac{1}{\sqrt{c}} \tag{1}$$

where $c = L/K$ is the compression rate. To explicitly model the epistemic uncertainty of the extracted memory skills, the PLMS outputs a mean vector μ_t and a log-variance vector $\log \sigma_t^2$ through two parallel linear layers applied to C_t. During training, to enable gradient backpropagation through the stochastic sampling process, we employ the reparameterization trick:

$$\hat{S}_{new} = \mu_t + \sigma_t \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{2}$$

To regularize the distribution, we apply a Kullback-Leibler (KL) divergence loss $\mathcal{L}_{KL}$ to push the approximate posterior towards a standard Gaussian prior.

Gated Latent Consolidator. To mitigate the semantic dilution problem and filter out useless chat noise (e.g., redundant greetings), the GLC strictly controls how much new information should be permanently stored. To explicitly quantify the novelty of incoming information against the established memory skills, we introduce an information gain gate $g_t \in (0,1)^{K\times D}$. This gate is learned via a multi-layer perceptron operating on the concatenated representations of the new and old memory skills:

$$g_t = \sigma\Big(W_2\text{ReLU}\big(W_1[\hat{S}_{new} \parallel S_{old}]\big)\Big) \tag{3}$$

To maintain the temporal sequence of the latent flow while gracefully forgetting outdated elements, we utilize a Gated Recurrent Unit (GRU) core. The latent memory skills are conditionally updated step-by-step:

$$S_t = g_t \odot \text{GRU}(\hat{S}_{new}, S_{old}) + (1 - g_t) \odot S_{t-1} \tag{4}$$

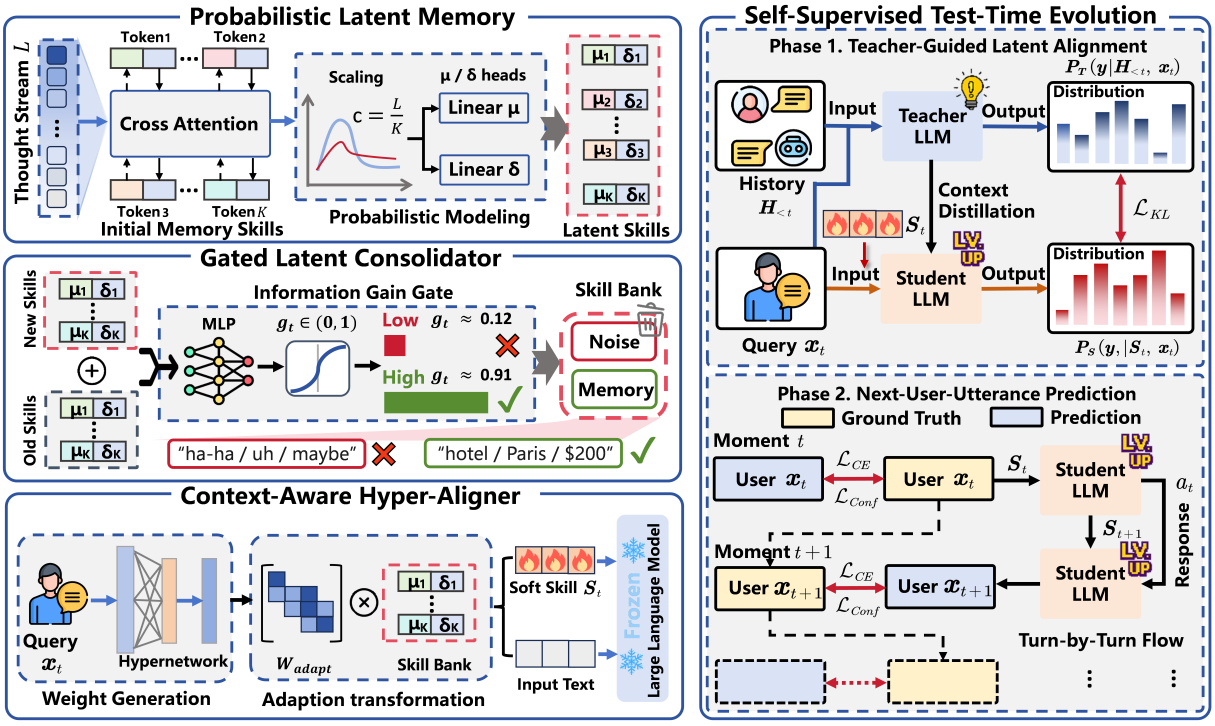

Figure 2: Illustration of our **ThinkFlow** (Left) and training strategy (Right). The architecture is an end-to-end memory system designed to extract, filter, and align historical information without relying on explicit text generation.

Context-Aware Hyper-Aligner. To bridge the distributional gap between historical latent skill representations and the current semantic context of the language model, the CAHA translates historical memory skills dynamically. To achieve this fine-grained adaptation without adding massive parameters, we adopt a hypernetwork approach. It generates a low-rank transformation matrix $W_{adapt} \in \mathbb{R}^{D\times D}$ based on the linguistic features of the current user query x_t:

$$W_{adapt} = W_{down}(x_t) \times W_{up}(x_t) \tag{5}$$

where $W_{down} \in \mathbb{R}^{D\times r}$ and $W_{up} \in \mathbb{R}^{r\times D}$ form a bottleneck of rank r. To compute the final aligned memory skills, we perform a residual projection: $S_{aligned} = S_t + S_t \times W_{adapt}$. Finally, to condition the generation process, these aligned skills are prepended to the language model as soft prompts.

3.2 Self-Supervised Test-Time Evolution

Traditional memory models suffer from a severe deployment bottleneck: they rely heavily on static, human-annotated dialogue datasets, making them incapable of adapting to individual user idiosyncrasies after deployment. To achieve true personalization without requiring manual labels, ThinkFlow introduces a Self-Supervised Test-Time Evolution paradigm. This two-phase training strategy allows the agent to continuously refine its latent memory skills purely from real-world interactions.

Phase 1: Teacher-Guided Latent Alignment. To address the cold-start problem during a user's first session, we conduct a Teacher-Guided Latent Alignment. To transfer the comprehensive contextual understanding from a computationally expensive text-based teacher to our efficient latent memory skill student, the powerful teacher model reads the full plaintext history $H_{<t}$ and outputs a target probability distribution $P_T(y \mid H_{<t}, x_t)$. The student model only reads the current query x_t and the latent memory skills S_t, outputting a predictive distribution $P_S(y \mid S_t, x_t)$. To force the student to mimic the teacher and establish the initial memory skill anchor, we minimize objective:

$$\mathcal{L}_{Phase1} = \mathcal{D}_{KL}\Big(P_T(y \mid H_{<t}, x_t) \parallel P_S(y \mid S_t, x_t)\Big) + \alpha\mathcal{L}_{KL} \tag{6}$$

Phase 2: Next-User-Utterance Prediction. To enable continuous lifelong learning without manual annotations, the teacher model goes offline from Session 2 onwards. Drawing from *Predictive Coding* in cognitive science, human brains constantly generate top-down predictions of future sensory inputs and update their internal models based on prediction errors. Inspired by this, we cast the memory skill evolution problem as a *Next-User-Utterance Prediction* task, utilizing the user's implicit feedback as an abundant and free supervisory signal. At turn t, the agent generates response a_t and updates the memory skills to S_{t+1}. At turn $t+1$, the user provides a new query x_{t+1}. To construct a pure self-supervised signal that grounds the memory skill evaluation on real-world predictive success, the system

uses the latent memory skills S_{t+1} and its previous response a_t to predict the user's next utterance x_{t+1}.

If the memory skills S_{t+1} fail to capture the core context, the model will fail to predict the user's reaction, resulting in a high cross-entropy loss $\mathcal{L}_{CE}$. To shield the memory skill bank from being polluted by bad updates, this high loss directly forces the GLC gate to close. Furthermore, to encourage the model to maintain sharp, confident predictions and prevent ambiguous semantic collapse, we introduce a minimum entropy confidence prior $\mathcal{L}_{Conf}$. The loss for Phase 2 is formulated as:

$$\begin{aligned}\mathcal{L}_{Phase2} = \mathcal{L}_{CE}(x_{t+1} \mid S_{t+1}, a_t) \\ + \beta\mathcal{L}_{Conf} + \alpha\mathcal{L}_{KL}\end{aligned} \tag{7}$$

4 Experiments

4.1 Experimental Settings

Datasets and Benchmarks. Following (Ong et al., 2025), we evaluate our method on three long-term open-domain conversation datasets: **Multi-Session Chat** (MSC) (Xu et al., 2022), **Conversation Chronicles** (CC), (Jang et al., 2023), and **GapChat** (GC) (Zhang et al., 2023) and a personalized memory question answering benchmark: **PersonaMem** (Jiang et al., 2025a). More details are shown in Appendix A.

Models and Baselines. For backbone, we evaluate on two open-source long-context LLMs: 1) **Llama-3.2 3B-Instruct** (Grattafiori et al., 2024). 2) **Qwen3-8B** (Yang et al., 2025). We compare our **ThinkFlow** against various baselines. 1) **Long Context**: which use all the conversation histories. 2) **Structured Retrieval-Augmented Generation**: **GraphRAG** (Edge et al., 2024) and **MemTree** (Rezazadeh et al., 2025). 3) **Agentic Memory Management**: **MemGPT** (Packer et al., 2023), **A-Mem** (Xu et al., 2025b), **Mem0** (Chhikara et al., 2025), and **MemoryOS** (Ong et al., 2025). 4) **Memory-Augmented Generation**: **MemoChat** (Lu et al., 2023), **MemoryBank** (Zhong et al., 2024), **Rsum** (Wang et al., 2025b), **COMEDY** (Chen et al., 2025), **LD-Agent** (Li et al., 2025a), **THEANINE** (Ong et al., 2025), and **LightMem** (Fang et al., 2026). More details are shown in Appendix B.

Evaluation Metrics. We comprehensively evaluate our method on automatic metrics and accuracy: 1) **Automatic Metrics.** On CC, MSC and GC, we use BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), BertScore (Zhang et al., 2019), and Mauve (Pillutla et al., 2021) to automatically evaluate turn-by-turn response generation. 2) **Accuracy.** On PersonaMem, performance evaluation is based on the accuracy of the generated responses, specifically the proportion of responses that correctly match the user's current persona and the conversational context.

Implementation Details. To maintain computational efficiency, we employ LoRA (Hu et al., 2022) on LLMs. For the core parameter, the default number of probabilistic latent memory skills is set to $K = 10$. To ensure strict alignment in the representation space and vocabulary during context distillation, both the Teacher and Student LLMs are instantiated from the same base LLMs. More details are shown in Appendix C.

4.2 Performance in Long-Term Generation

Bypassing the explicit text bottleneck via latent memory is crucial for preventing information loss in continuous dialogues. Table 1 presents the evaluation results for turn-by-turn generation. Directly feeding the full dialogue history into the long context yields sub-optimal performance, highlighting the severe issue of semantic dilution. Furthermore, while various explicit memory paradigms improve performance over the long-context baseline, their gains are fundamentally constrained by the information bottleneck of discrete text summaries. In contrast, our proposed ThinkFlow framework consistently achieves the best overall performance across all datasets. Notably, ThinkFlow demonstrates massive improvements in the Mauve metric, indicating much more coherent and human-like responses without losing subtle conversational nuances.

4.3 Performance in Personalized Memory

Tracking the dynamic evolution of user states requires continuous latent updating rather than static text retrieval. Table 2 shows that ThinkFlow consistently achieves the highest average accuracy among comparable methods on the PersonaMem benchmark across extremely long contexts (up to 1M tokens), particularly excelling in complex, dynamic scenarios like tracking user evolution. Remarkably, our ThinkFlow-8B rivals or even surpasses current powerful closed-source and massive open-source (405B) models. These results validate that maintaining memory purely in the latent space effectively overcomes the text bottleneck, accurately capturing user states for lifelong companionship.

4.4 Ablation Study

Both structural feature disentanglement and continuous self-supervised evolution are indispensable for maintaining high-quality companionship. Table 3 presents the ablation study of ThinkFlow on the Qwen3-8B backbone. Removing core structural modules (*w/o* PLMS, GLC, CAHA) leads to substantial performance degradation across datasets, particularly in the Mauve metric. The severe drop observed when removing PLMS validates the fundamental necessity of compressing conversational flow into disentangled skill vectors. Furthermore, omitting the two training phases (*w/o* Phase1, Phase2) also results in noticeable declines, confirming that both the initial teacher-guided alignment and the continuous test-time evolution are critical for achieving dynamic, label-free personalization.

Backbone	Methods	CC				MSC				GC			
		B-4	R-L	Bert	Mauve	B-4	R-L	Bert	Mauve	B-4	R-L	Bert	Mauve
Llama3.2-3B	Long Context (128K)	0.54	10.09	43.12	35.53	0.47	9.94	44.61	45.34	0.20	3.79	33.12	18.85
	Explicit Memory Paradigm												
	GraphRAG (2024)	0.81	11.37	45.19	48.23	0.62	10.15	45.03	47.61	0.31	4.57	34.11	19.23
	MemTree (2025)	0.73	11.22	44.57	51.78	0.56	10.02	44.51	49.46	0.28	4.42	33.58	18.65
	MemGPT (2023)	0.75	11.53	40.18	37.52	0.52	9.27	40.65	38.81	0.27	4.31	33.17	18.53
	A-Mem (2025b)	0.64	10.35	44.62	31.28	0.47	10.05	44.51	42.48	0.25	4.19	33.43	17.89
	Mem0 (2025)	0.67	10.53	43.15	31.49	0.46	9.72	43.24	44.89	0.26	4.23	33.51	18.11
	MemoryOS (2025)	0.69	10.61	43.47	32.17	0.48	9.83	43.69	45.13	0.27	4.37	33.73	18.41
	MemoChat (2023)	0.83	11.91	40.73	38.45	0.55	10.11	41.27	39.51	0.26	4.13	32.95	17.93
	MemoryBank (2024)	1.06	13.22	42.10	44.30	0.63	11.27	42.97	42.02	0.33	5.17	34.61	19.87
	Rsum (2025b)	1.01	12.87	41.83	43.19	0.61	11.03	42.59	41.77	0.31	5.09	34.27	19.43
	COMEDY (2025)	0.61	9.83	38.97	35.21	0.43	8.71	39.81	36.43	0.19	3.51	31.47	16.53
	LD-Agent (2025a)	1.13	13.51	42.83	45.67	0.67	11.53	43.57	43.11	0.35	5.31	35.13	20.17
	THEANINE (2025)	1.09	13.11	42.41	56.73	0.65	11.19	43.19	51.27	0.34	5.23	34.87	21.39
	LightMem (2026)	0.70	11.68	45.21	54.47	0.65	11.80	**46.02**	58.38	0.29	4.96	35.04	19.05
	Latent Memory Paradigm												
	SoftCoT (2025c)	0.25	3.68	27.27	23.97	0.12	2.81	26.30	37.06	0.06	0.84	22.42	31.15
	Co-processor (2025)	0.88	12.05	43.09	**63.12**	0.52	9.54	43.16	**58.91**	0.28	4.41	32.71	**42.01**
	MemGen (2026a)	0.62	9.98	42.16	41.56	0.46	8.70	42.42	49.11	0.29	4.63	32.78	26.85
	ThinkFlow (Ours)	**1.21**	**14.84**	**45.62**	58.85	**0.92**	**13.21**	45.77	51.81	**0.77**	**9.03**	**38.89**	39.56
Qwen3-8B	Long Context (128K)	0.66	10.72	45.89	21.43	0.41	9.20	43.85	46.49	0.43	7.51	33.91	21.13
	Explicit Memory Paradigm												
	GraphRAG (2024)	1.15	13.21	46.13	34.17	0.65	11.13	45.17	42.51	0.54	8.37	35.21	24.53
	MemTree (2025)	1.09	12.97	45.97	31.59	0.59	10.68	44.49	41.76	0.51	8.10	34.97	23.85
	MemGPT (2023)	1.06	12.82	45.41	39.68	0.81	11.42	44.68	46.24	0.52	8.23	34.83	23.91
	A-Mem (2025b)	0.94	13.63	43.79	26.85	0.77	12.31	44.62	44.15	0.49	8.07	34.59	22.87
	Mem0 (2025)	0.72	12.23	43.12	45.89	0.63	11.57	43.32	45.72	0.51	8.11	34.67	23.41
	MemoryOS (2025)	0.81	12.45	43.51	46.23	0.67	11.73	43.83	46.11	0.53	8.19	34.73	23.77
	MemoChat (2023)	0.95	13.73	42.19	42.13	0.77	12.11	43.91	44.53	0.47	7.83	33.87	22.19
	MemoryBank (2024)	1.22	14.91	43.40	47.53	0.88	13.26	45.14	46.33	0.61	8.93	35.53	25.13
	Rsum (2025b)	1.17	14.53	43.13	46.87	0.85	13.01	44.87	45.89	0.58	8.71	35.29	24.83
	COMEDY (2025)	0.83	11.51	40.53	38.67	0.61	10.37	41.97	41.23	0.41	6.77	32.53	20.31
	LD-Agent (2025a)	1.27	15.13	43.73	48.21	0.91	13.51	45.43	47.19	0.63	9.17	35.81	25.67
	THEANINE (2025)	1.23	14.77	43.27	58.19	0.89	13.19	45.11	53.47	0.62	8.99	35.69	26.83
	LightMem (2026)	1.05	13.53	46.43	32.31	0.63	11.44	45.35	49.26	0.38	6.73	34.23	30.22
	Latent Memory Paradigm												
	SoftCoT (2025c)	0.23	4.92	34.28	20.07	0.25	5.48	36.48	41.57	0.07	1.53	26.04	38.29
	Co-processor (2025)	0.59	9.63	42.55	37.43	0.53	9.49	42.45	59.74	0.19	3.74	32.45	49.96
	MemGen (2026a)	0.64	9.86	41.51	26.92	0.43	8.66	41.66	59.80	0.22	4.16	31.42	43.18
	ThinkFlow (Ours)	**2.18**	**17.83**	**47.77**	**77.20**	**1.12**	**14.33**	**46.72**	**66.48**	**1.10**	**14.21**	**44.82**	**65.26**

Table 1: Evaluation of generation performance (%) on conversational datasets. The **best** result is bolded and the second-best is underlined. *B-4 = BLEU-4, R-L = ROUGE-L, and Bert = BertScore.

4.5 Framework Analysis

The latent memory naturally disentangles static facts from continuously shifting user states, forming interpretable manifolds. Figure 3 visualizes the latent memory space using t-SNE on the PersonaMem benchmark. It reveals a clear structural separation among different cognitive tasks. Notably, representations for "Shared Facts" form a dense, highly localized cluster, reflecting the static nature of objective knowledge. In stark contrast, the manifold for "Track Evolution" spans a wide, continuous region. This visually confirms that our framework successfully tracks shifting user states within a fluid latent space, rather than relying on rigid text contents and summaries.

Expanding the capacity of latent cognitive receptors enables finer-grained feature disentanglement, thereby consistently enhancing personalization. Figure 4 illustrates the impact of scaling the number of latent memory skills (K) on the PersonaMem benchmark with a 1M context. As K increases from 1 to 10, the average accuracy exhibits a robust upward trend. Since the benchmark involves 7 task categories, scaling K to 10 acts as an upper bound to test performance when memory capacity exceeds task complexity. This

Length	Methods	Revisit Reasons	Latest Prefs	Shared Facts	Track Evolution	New Ideas	New Scenarios	Aligned Recs	Average
32K	*Closed-source LLMs*								
	GPT-4o-mini*	74.00	18.00	29.00	48.00	16.00	7.00	29.00	31.57
	Comparable Methods (Qwen3-8B)								
	MemTree	14.14	17.06	20.93	15.11	9.68	14.04	27.27	16.89
	A-Mem	13.03	11.76	13.10	12.16	12.15	13.51	13.64	12.76
	MemoryBank	63.64	23.53	51.94	56.12	11.83	29.82	**36.36**	39.03
	MemGen	65.66	17.65	46.51	64.75	12.90	24.56	32.73	37.82
	ThinkFlow-8B (Ours)	**67.68**$^{+2.02}$	**29.41**$^{+5.88}$	**55.81**$^{+3.87}$	**70.50**$^{+5.75}$	**21.51**$^{+8.61}$	**31.58**$^{+1.76}$	34.55$^{-1.81}$	**44.43**$^{+5.40}$
128K	*Closed- & Open-source LLMs*								
	GPT-4o-mini*	70.00	34.00	55.00	60.00	10.00	33.00	41.00	43.29
	Claude3.7-Sonnet*	57.00	9.00	25.00	45.00	28.00	29.00	20.00	30.43
	Llama3.1-405B*	41.00	31.00	38.00	38.00	20.00	21.00	37.00	32.29
	Comparable Methods (Qwen3-8B)								
	MemTree	16.73	22.63	16.37	17.60	18.73	23.47	24.93	20.07
	A-Mem	12.64	9.47	8.19	9.38	9.07	7.98	11.75	9.78
	MemoryBank	53.16	38.11	26.32	57.77	15.06	24.88	30.95	35.18
	MemGen	57.25	40.65	22.22	61.58	19.11	26.76	31.81	37.05
	ThinkFlow-8B (Ours)	**63.20**$^{+5.95}$	**50.92**$^{+10.27}$	**29.24**$^{+2.92}$	**63.64**$^{+2.06}$	**20.08**$^{+0.97}$	**27.23**$^{+0.47}$	**32.09**$^{+0.28}$	**40.91**$^{+3.86}$
1M	*Closed-source LLMs*								
	Gemini2.0-Flash*	68.00	39.00	42.00	62.00	12.00	49.00	41.00	44.71
	Comparable Methods (Qwen3-8B)								
	MemTree	11.91	22.79	20.83	11.11	**20.08**	25.08	22.86	19.24
	A-Mem	9.79	6.64	8.33	7.11	5.23	5.76	6.79	7.09
	MemoryBank	51.49	33.46	34.03	61.33	19.12	29.15	26.78	36.48
	MemGen	52.77	39.32	35.42	**63.11**	19.53	32.20	29.64	38.86
	ThinkFlow-8B (Ours)	**68.09**$^{+15.32}$	**40.76**$^{+1.44}$	**38.89**$^{+3.47}$	62.22$^{-0.89}$	18.98$^{-1.10}$	**33.22**$^{+1.02}$	**31.43**$^{+1.79}$	**41.94**$^{+2.36}$

Table 2: Accuracy on the PersonaMem benchmark. For comparable methods, the **best** result is bolded and the second-best is underlined. Baseline results marked with * are retrieved from (Jiang et al., 2025a). Gain ($^{+\cdot}$) and loss ($^{-\cdot}$) represent the relative performance of our **ThinkFlow-8B** compared to the best comparable method in each task.

Methods	B-4	R-L	Bert	Mauve
ThinkFlow (Ours)	**2.18**	**17.83**	**47.77**	**77.20**
w/o PLMS	1.88	16.54	45.53	66.48
w/o GLC	1.94	16.99	45.59	64.98
w/o CAHA	2.04	17.45	44.98	67.65
w/o Phase1	2.00	17.33	44.84	67.87
w/o Phase2	2.04	17.27	44.87	69.02

Table 3: Ablation study of our method (Qwen3-8B) On CC dataset. More results are show in Appendix D.

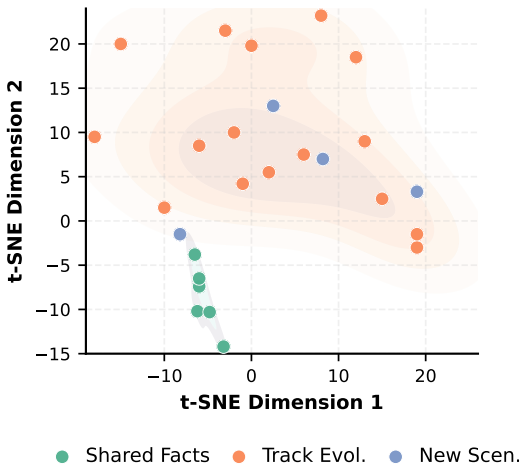

Figure 3: t-SNE visualization of latent memory across three typical tasks on PersonaMem.

improvement is particularly pronounced in complex reasoning sub-tasks such as "Revisit Reasons" and "Track Evolution". It demonstrates that allocating more latent skill vectors allows the model to capture and disentangle a richer set of user nuances, effectively boosting the resolution of the memory space without context windows.

Operating exclusively in the latent space eliminates the massive token overhead and computational latency inherent in explicit memory pipelines. Figure 5 illustrates the token-time scaling efficiency across interaction sessions. ThinkFlow demonstrates remarkable efficiency, consuming the least tokens (2,857) and requiring the shortest time (12.1s). In contrast, explicit methods like MemTree and A-Mem suffer from severe computational bottlenecks, generating tens of thousands of hidden tokens and taking nearly 8 times longer to execute. This proves that our latent memory paradigm is not

Dialogue History & Context	*[Relationship: Husband & Wife]*
Session 3 *(Past Preference)*	**User**: I saw this amazing jacket in the window display today... I'm willing to spend the money to look good.
Session 5 *(Relationship Crisis)*	**User**: I can't believe you've been seeing someone else behind my back! ... I feel like everything we had has been shattered.
Current Turn *(Preference Shift)*	**User**: *"That seems like a lifetime ago. I don't feel much like dressing up these days."*

Baseline: MemTree (Explicit Retrieval)	**Ours: ThinkFlow** (Probabilistic Latent)
MEMORY STATUS *Retrieved Nodes (Similarity triggered):* • [Node 49]: I saw this amazing jacket... • [Node 82]: loved trying out new styles...	**MEMORY STATUS** *Latent State Evolution:* Distribution transitions smoothly from `[Interest in Fashion]` → `[Deep Sadness]`.
AGENT RESPONSE I understand where you're coming from... **maybe we could try something new together, like exploring a new style!**	**AGENT RESPONSE** I completely understand. **When you're dealing with so much pain, it's completely normal to lose interest in things like dressing up.** I'm here for you, take all the time you need.
ANALYSIS × **Failure:** Trapped by the surface fact "dressing up". Rigidly retrieves outdated explicit history and entirely misses the user's emotional collapse.	**ANALYSIS** ✓ **Success:** Successfully tracks the continuous latent emotion, naturally adapting to the loss of interest without rigid similarity dependency.

Table 4: Case study on tracking user evolution on CC dataset.

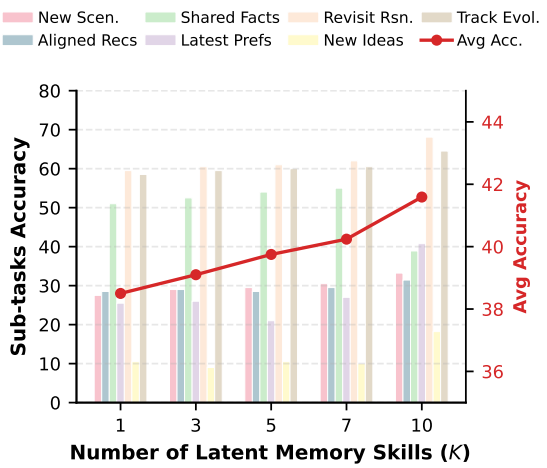

Figure 4: Latent memory scaling on PersonaMem with 1M context. More results are shown in Appendix F.

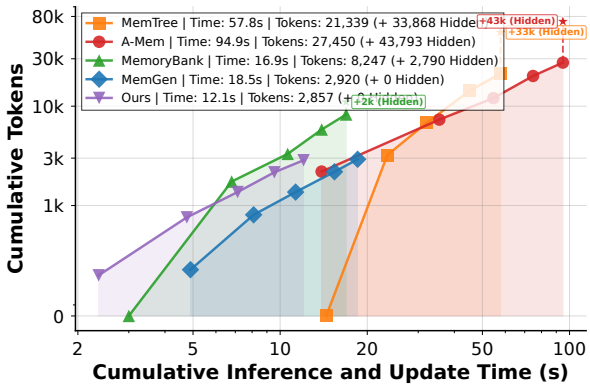

Figure 5: Token-time scaling efficiency on CC dataset, where "Hidden" tokens refer to the intermediate memory processed internally by explicit baselines.

only highly accurate but also exceptionally lightweight and scalable for real-time companionship.

4.6 Case Study

Explicit retrieval often traps models in outdated contexts, whereas latent memory smoothly adapts to profound emotional shifts. Table 4 illustrates a multi-session dialogue where a severe relationship crisis alters the user's past interests. The explicit baseline MemTree is easily trapped by the surface fact "dressing up", rigidly retrieving outdated history and delivering an emotionally tone-deaf response. In contrast, ThinkFlow dynamically tracks the smooth transition of the latent emotional state from initial interest to deep sadness. Consequently, our method generates a highly empathetic and contextually appropriate response without relying on similarity retrieval.

5 Conclusions

In this work, we propose ThinkFlow, a novel end-to-end latent memory framework designed to empower lifelong conversational agents. By departing from traditional explicit textual memory pipelines that suffer from severe information bottlenecks, ThinkFlow dynamically compresses continuous conversational flows into disentangled probabilistic memory skills. Furthermore, our introduced self-supervised test-time evolution paradigm effectively overcomes the static deployment

bottleneck, allowing the agent to continuously adapt to evolving user nuances through implicit predictive feedback without requiring manual annotations. Extensive experiments across multiple long-term open-domain conversation datasets and the PersonaMem benchmark demonstrate that ThinkFlow not only significantly outperforms state-of-the-art explicit memory baselines in personalization and response quality, but also achieves remarkable token and temporal efficiency.

Limitations

While ThinkFlow establishes a highly efficient and self-evolving paradigm for lifelong memory, it opens up several ambitious avenues for future exploration that bound our current scope. First, due to the extraordinary token efficiency and minimal latency of our probabilistic latent memory skills, the current framework processes ultra-long multi-session histories so effortlessly on standard hardware that we have not yet established its upper breakdown limits. Pushing the boundary of this architecture to extreme-scale, life-long contexts—such as infinite multi-modal continuous streams involving lifelong visual and audio inputs—remains an untapped territory, as our continuous latent space is naturally suited for cross-modal fusion rather than being restricted to text.

Second, our self-supervised test-time evolution currently utilizes the "next-user-utterance prediction" as its primary supervisory signal, which already yields robust results. However, given the strong feature disentanglement capabilities observed in our latent manifolds, this mechanism is currently underutilized. Future work should elevate this predictive coding mechanism to anticipate highly complex, long-horizon user behavioral trajectories or to simulate intricate multi-agent societal dynamics. Consequently, our current limitation primarily lies in the fact that the framework's full cognitive potential has yet to be unleashed across broader, cross-modal, and societal-scale multi-agent scenarios.

Acknowledgements

This work was supported by the National Natural Science Foundation of China 62576120 and the Major Key Project of PCL2025A11 and PCL2024A08. Thanks for the support provided by OpenI Community (https://openi.pcl.ac.cn).

References

Ian A Apperly and Stephen A Butterfill. 2009. Do humans have two systems to track beliefs and belief-like states? *Psychological review*, 116(4):953.

Sanghwan Bae, Donghyun Kwak, Soyoung Kang, Min Young Lee, Sungdong Kim, Yuin Jeong, Hyeri Kim, Sang-Woo Lee, Woomyoung Park, and Nako Sung. 2022. Keep me updated! memory management in long-term conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3769–3787.

Nuo Chen, Hongguang Li, Jianhui Chang, Juhua Huang, Baoyuan Wang, and Jia Li. 2025. Compress to impress: Unleashing the potential of compressive memory in real-world long-term conversations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 755–773.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*.

Andy Clark. 2013. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, Huajun Chen, and Ningyu Zhang. 2026. Lightmem: Lightweight and efficient memory-augmented generation. In *The Fourteenth International Conference on Learning Representations*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, et al. 2025. Memory in the age of ai agents. *arXiv preprint arXiv:2512.13564*.

Jihyoung Jang, Minseong Boo, and Hyounghun Kim. 2023. Conversation chronicles: Towards diverse temporal and relational dynamics in multi-session conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13584–13606.

Jihyoung Jang, Taeyoung Kim, and Hyounghun Kim. 2024. Mixed-session conversation with egocentric memory. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11786–11815.

Bowen Jiang, Zhuoqun Hao, Young Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo Jose Taylor, and Dan Roth. 2025a. Know me, respond to me: Benchmarking LLMs for dynamic user profiling and personalized responses at scale. In *Second Conference on Language Modeling*.

Bowen Jiang, Yuan Yuan, Maohao Shen, Zhuoqun Hao, Zhangchen Xu, Zichen Chen, Ziyi Liu, Anvesh Rao Vijjini, Jiashu He, Hanchao Yu, et al. 2025b. Personamem-v2: Towards personalized intelligence via learning implicit user personas and agentic memory. *arXiv preprint arXiv:2512.06688*.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. 2025. Memory OS of AI agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25972–25981. Association for Computational Linguistics.

Cai Ke, Yiming Du, Bin Liang, Yifan Xiang, Lin Gui, Zhongyang Li, Baojun Wang, Yue Yu, Hui Wang, Kam-Fai Wong, et al. 2025. Flexibly utilize memory for long-term conversation via a fragment-then-compose framework. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21130–21147.

Cai Ke, Bin Liang, Xin Liu, Yue Yu, Hui Wang, and Ruifeng Xu. 2026a. Dynamic memory forest: Constructing and tracing conversational trajectories for long-term conversation. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 767–777.

Cai Ke, Jiangyue Yan, Han Zhang, Liu Xin, Zike Yuan, Yue Yu, Hui Wang, and Ruifeng Xu. 2026b. Interactive memory learning for long-term conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2026*.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2026. LLMs get lost in multi-turn conversation. In *The Fourteenth International Conference on Learning Representations*.

Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. 2025a. Hello again! LLM-powered personalized agent for long-term dialogue. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5259–5276.

Hengli Li, Chenxi Li, Tong Wu, Xuekai Zhu, Yuxuan Wang, Zhaoxin Yu, Eric Hanchen Jiang, Song-Chun Zhu, Zixia Jia, Ying Nian Wu, et al. 2025b. Seek in the dark: Reasoning via test-time instance-level policy gradient in latent space. *arXiv preprint arXiv:2505.13308*.

Bin Liang, Cai Ke, Runcong Zhao, Qinglin Zhu, Lin Gui, Yue Yu, Hui Wang, Ruifeng Xu, and Kam-Fai Wong. 2026. Meta-memory for large language models. *IEEE Transactions on Audio, Speech and Language Processing*.

CY Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, Barcelona, Spain*, pages 74–81.

Luyang Liu, Jonas Pfeiffer, Jiaxing Wu, Jun Xie, and Arthur Szlam. 2025. Deliberation in latent space via differentiable cache augmentation. In *Forty-second International Conference on Machine Learning*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024a. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Zhiwei Liu, Weiran Yao, Jianguo Zhang, Liangwei Yang, Zuxin Liu, Juntao Tan, Prafulla K Choubey, Tian Lan, Jason Wu, Huan Wang, et al. 2024b. Agentlite: A lightweight library for building and advancing task-oriented llm agent system. *arXiv preprint arXiv:2402.15538*.

Junru Lu, Siyu An, Mingbao Lin, Gabriele Pergola, Yulan He, Di Yin, Xing Sun, and Yunsheng Wu. 2023. Memochat: Tuning llms to use memos for consistent long-range open-domain conversation. *arXiv preprint arXiv:2308.08239*.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of llm agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870.

Kai Mei, Xi Zhu, Wujiang Xu, Wenyue Hua, Mingyu Jin, Zelong Li, Shuyuan Xu, Ruosong Ye, Yingqiang Ge, and Yongfeng Zhang. 2024. Aios: Llm agent operating system. *arXiv preprint arXiv:2403.16971*.

Beren Millidge, Anil Seth, and Christopher L Buckley. 2021. Predictive coding: a theoretical and experimental review. *arXiv preprint arXiv:2107.12979*.

Kai Tzu-iunn Ong, Namyoung Kim, Minju Gwak, Hyungjoo Chae, Taeyoon Kwon, Yohan Jo, Seungwon Hwang, Dongha Lee, and Jinyoung Yeo. 2025. Towards lifelong dialogue agents via timeline-based memory management. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8631–8661.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G Patil, Ion Stoica, and Joseph E Gonzalez. 2023. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828.

Alireza Rezazadeh, Zichao Li, Wei Wei, and Yujia Bao. 2025. From isolated conversations to hierarchical schemas: Dynamic tree memory representation for LLMs. In *The Thirteenth International Conference on Learning Representations*.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. 2024. Raptor: Recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations*.

Dana Schneider, Rebecca Lam, Andrew P Bayliss, and Paul E Dux. 2012. Cognitive load disrupts implicit theory-of-mind processing. *Psychological science*, 23(8):842–847.

Juntao Tan, Liangwei Yang, Zuxin Liu, Zhiwei Liu, Rithesh RN, Tulika Manoj Awalgaonkar, Jianguo Zhang, Weiran Yao, Ming Zhu, Shirley Kokane, et al. 2025. Personabench: Evaluating ai models on understanding personal information through accessing (synthetic) private user data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 878–893.

Piaohong Wang, Motong Tian, Jiaxian Li, Yuan Liang, Yuqing Wang, Qianben Chen, Tiannan Wang, Zhicong Lu, Jiawei Ma, Yuchen Eleanor Jiang, et al. 2025a. O-mem: Omni memory system for personalized, long horizon, self-evolving agents. *arXiv e-prints*, pages arXiv–2511.

Qingyue Wang, Yanhe Fu, Yanan Cao, Shuai Wang, Zhiliang Tian, and Liang Ding. 2025b. Recursively summarizing enables long-term dialogue memory in large language models. *Neurocomputing*, page 130193.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2025. Longmemeval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations*.

Derong Xu, Yi Wen, Pengyue Jia, Yingyi Zhang, Yichao Wang, Huifeng Guo, Ruiming Tang, Xiangyu Zhao, Enhong Chen, Tong Xu, et al. 2025a. Towards multi-granularity memory association and selection for long-term conversational agents. *arXiv preprint arXiv:2505.19549*.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025b. A-mem: Agentic memory for llm agents. In *Advances in Neural Information Processing Systems*.

Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. 2025c. Softcot: Soft chain-of-thought for efficient reasoning with llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23336–23351.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Xinlei Yu, Zhangquan Chen, Yongbo He, Tianyu Fu, Cheng Yang, Chengming Xu, Yue Ma, Xiaobin Hu, Zhe Cao, Jie Xu, et al. 2026. The latent space: Foundation, evolution, mechanism, ability, and outlook. *arXiv preprint arXiv:2604.02029*.

Guibin Zhang, Muxin Fu, and Shuicheng YAN. 2026a. Memgen: Weaving generative latent memory for self-evolving agents. In *The Fourteenth International Conference on Learning Representations*.

Qiang Zhang, Jason Naradowsky, and Yusuke Miyao. 2023. Mind the gap between conversations for improved long-term dialogue generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10735–10762.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yingyi Zhang, Junyi Li, Wenlin Zhang, Pengyue Jia, Xianneng Li, Yichao Wang, Derong Xu, Yi Wen, Huifeng Guo, Yong Liu, and Xiangyu Zhao. 2026b. Evoking user memory: Personalizing LLM via recollection-familiarity adaptive retrieval. In *The Fourteenth International Conference on Learning Representations*.

Zeyu Zhang, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2025. A survey on the memory mechanism of large language model-based agents. *ACM Transactions on Information Systems*, 43(6):1–47.

Zeyu Zhang, Rui Li, Xiaoyan Zhao, Yang Zhang, Wenjie Wang, Xu Chen, and Tat-Seng Chua. 2026c. Nextmem: Towards latent factual memory for llm-based agents. *arXiv preprint arXiv:2603.15634*.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19724–19731.

A Dataset Information

Long-Term Open-Domain Conversation. To comprehensively assess the model's proficiency in turn-by-turn response generation during prolonged interactions, we utilize three representative long-term open-domain conversation datasets: **Conversation Chronicles** (CC) (Jang et al., 2023), **Multi-Session Chat** (MSC) (Xu et al., 2022), and **GapChat** (GC) (Zhang et al., 2023): The primary motivation for selecting these

Datasets	# of Sessions	# of Episodes	# of Turns	Avg. Turns per Session	Avg. Turns per Episode
CC	1M	200K	11.7M	11.70	58.50
MSC	16K	5K	214K	13.38	42.80
GC	2.65K	0.65K	28.13K	10.62	43.28

Table 5: The statistics of CC, MSC and GC datasets.

Characteristic	32K Corpus	128K Corpus	1M Corpus
Basic Statistics			
Total Evaluation Samples	589	2,727	2,674
Avg. Query Length (Tokens)	464.6	416.3	415.1
Avg. Memory Length (Tokens)	24,193.2	151,851.0	911,733.4
Distribution by Evaluated Skill (Query Type)			
Recall Shared Facts	129	171	144
Acknowledge Latest Prefs	17	866	768
Track Evolution	139	341	225
Revisit Reasons	99	269	235
Aligned Recommendations	55	349	280
Suggest New Ideas	93	518	727
New Scenarios	57	213	295

Table 6: The statistics of PersonaMem benchmark across different memory corpus scales.

specific datasets lies in their human-centric construction methodology. They are meticulously curated through extensive crowdsourcing involving real human participants, ensuring that the dialogue flows, topic transitions, and relationship developments accurately mirror the complexity of real-world open-domain social dynamics. Consequently, evaluating on these datasets is essential for verifying whether Large Language Models (LLMs) can generate natural, coherent, and highly engaging responses that genuinely align with human expectations in lifelong conversational scenarios. Specifically, **CC** focuses on temporal dynamics and intricate relationship evolutions over extensive sessions. **MSC** provides a massive corpus of authentic multi-session chats where speakers incrementally learn and recall each other's historical interests. **GC** introduces the critical dimension of realistic time gaps (ranging from minutes to years) between sessions, challenging the agent to emulate a human-like perception of the passage of time.

Personalized Memory Question Answering. To further evaluate the crucial role of personalized memory—specifically, the model's capability to accurately track and reason over dynamic user states across varying context lengths—we use the PersonaMem benchmark (Jiang et al., 2025a). While open-domain generation tests general conversational fluency, maintaining true lifelong companionship fundamentally relies on this personalized memory tracking. This benchmark includes about 15 real-world interaction scenarios (such as medical advice, travel planning, and various recommendations), simulating dynamic, long-term conversations between users and AI assistants. As shown in Table 6, we divide the benchmark into three scales—32K, 128K, and 1M—to test the model's retrieval and reasoning limits. As the memory size grows, the average dialogue history expands dramatically from about 24,000 tokens to over 910,000 tokens, while the user's query length stays short at around 400 tokens. This short query, ultra-long memory" setup strongly challenges the model's ability to find precise information amidst noise. Furthermore, the benchmark tests 7 specific skills. As the table shows, these tasks range from simple fact retrieval (e.g., Recall Shared Facts) to complex reasoning tasks that require tracking changing preferences (Track Evolution) or generalizing to entirely new situations (Suggest New Ideas/New Scenarios). This diverse mix of tasks ensures we evaluate the model's true understanding" of a user profile, rather than memorizing long texts.

B Compared Baselines

To systematically evaluate the effectiveness of ThinkFlow, we compare it against a diverse set of state-of-the-art baselines. Following the taxonomy established in our Related Work, these baselines are categorized into the Explicit Memory Paradigm and the Latent Memory Paradigm.

B.1 Explicit Memory Paradigm

This paradigm predominantly relies on discrete, human-readable text to construct and manage historical interactions. Based on their core mechanisms, we further divide them into three sub-categories:

B.1.1 Structured Retrieval-Augmented Generation

These methods focus on organizing past conversational logs into structured topologies, such as graphs or trees,

to facilitate complex context retrieval.

- **GraphRAG** (Edge et al., 2024): GraphRAG builds an entity knowledge graph from the source text and groups related entities into hierarchical community summaries. When answering a query, it retrieves information from these structured summaries, making it particularly effective at answering global questions that require understanding the entire document corpus.
- **MemTree** (Rezazadeh et al., 2025): MemTree organizes conversational history into a dynamic, tree-structured hierarchy. Each node within the tree stores aggregated text and its semantic embedding at varying levels of abstraction. It continuously updates this structure by comparing new information with existing nodes, enabling the agent to handle complex reasoning tasks over extended contexts.

B.1.2 Agentic Memory Management

Approaches in this category utilize autonomous agent architectures or multi-tiered storage systems to actively read, write, and govern memory blocks.

- **MemGPT** (Packer et al., 2023): Inspired by traditional operating systems, MemGPT introduces a virtual context management system. It creates the illusion of an infinite context window by intelligently paging relevant historical data between a large external storage disk and the LLM's limited working memory.
- **Mem0** (Chhikara et al., 2025): Mem0 operates through a two-step pipeline consisting of extraction and updating. It first identifies essential facts from conversational turns and subsequently employs tool-calling mechanisms to autonomously determine whether to insert, modify, remove, or entirely disregard the incoming information.
- **A-Mem** (Xu et al., 2025b): Drawing inspiration from the Zettelkasten note-taking method, A-Mem converts conversational interactions into atomic memory notes enriched with specific tags and keywords. It then utilizes embedding-based similarity searches to precisely retrieve relevant historical contexts.
- **MemoryOS** (Kang et al., 2025): Emulating a computer operating system, MemoryOS organizes interaction data across short-term, mid-term, and long-term storage tiers. It efficiently manages memory by clustering topic-related dialogues into structured segments and individual pages.

B.1.3 Memory-Augmented Generation

These frameworks are designed to dynamically extract discrete factual points or generate condensed textual summaries to build plug-and-play historical memory banks.

- **MemoChat** (Lu et al., 2023): MemoChat utilizes instruction-tuning to train language models to generate and consult internal "memos." It relies on a structured cycle of memorization, retrieval, and response generation to maintain logical consistency during extended conversations.
- **MemoryBank** (Zhong et al., 2024): Inspired by human cognitive processes, MemoryBank continuously archives conversation logs and distills them into hierarchical event summaries. This progressive summarization enables the agent to gradually align with the user's specific personality.
- **Rsum** (Wang et al., 2025b): Rsum introduces a recursive summarization framework. It sequentially merges the previously stored memory state with the latest dialogue segments to produce an updated summary, thereby preserving temporal continuity.
- **COMEDY** (Chen et al., 2025): Moving away from conventional external retrieval modules, COMEDY employs a unified model to concurrently handle memory generation, compression, and response formulation. It condenses user dynamics and historical events into a highly compact textual representation.
- **LD-Agent** (Li et al., 2025a): LD-Agent addresses long-term personalization through a decoupled, modular architecture. It distinctly isolates the independent processes of perceiving events, extracting user personas, and generating the final responses.
- **THEANINE** (Ong et al., 2025): In contrast to systems that routinely overwrite or discard obsolete data, THEANINE deliberately preserves all historical records. This inclusive strategy ensures that shifting user habits and outdated—yet contextually valuable—behavioral patterns are not permanently lost.
- **LightMem** (Fang et al., 2026): Inspired by the Atkinson-Shiffrin cognitive model, LightMem organizes interactions into a three-stage memory architecture: sensory, short-term, and long-term memory. It rapidly filters irrelevant information and consolidates interactions into topic-based groups, significantly reducing computational overhead while maintaining effective historical retrieval.

B.2 Latent Memory Paradigm

Unlike explicit methods constrained by the textual bottleneck, these approaches encode multi-session experiences directly into continuous vector spaces to support implicit agentic reasoning.

- **SoftCoT** (Xu et al., 2025c): SoftCoT utilizes a lightweight assistant model to generate instance-specific "soft thought tokens" in a continuous space.

These latent representations are then mapped into the primary language model's representation space via a trainable projection module, enabling intermediate reasoning without altering the main model's parameters.

- **Co-processor** (Liu et al., 2025): Co-processor augments a frozen language model with an offline, asynchronous module that operates directly on the model's key-value (KV) cache. It distills additional computational steps into latent embeddings within the KV-cache, improving the fidelity of subsequent decoding without generating discrete text tokens.

- **MemGen** (Zhang et al., 2026a): MemGen introduces a dynamic generative memory framework featuring a memory trigger and a memory weaver. It monitors the agent's current state to dynamically construct latent token sequences, functioning as machine-native memory that seamlessly interweaves continuous memory representations with the reasoning process.

C Implementation Details

As showin Table 7, LoRA targets the query, key, value, and output projection layers (`q_proj`, `k_proj`, `v_proj`, `o_proj`) with rank $r = 8$ and $\alpha = 16$. Within the memory components, the cross-attention module in the Probabilistic Latent Memory Skills (PLMS) utilizes 8 attention heads. In the Gated Latent Consolidator (GLC), the intermediate projection size of the MLP gate is set to half of the model's hidden dimension ($D/2$). For the Context-Aware Hyper-Aligner (CAHA), the bottleneck rank is set to $r = 8$.

During training, we optimize the framework using the AdamW optimizer with a learning rate of 2×10^{-5} and apply gradient clipping with a maximum norm of 1.0. For the loss constraints, the KL divergence regularization weight α is set to 0.1, and the minimum entropy confidence prior weight β for Phase 2 is also set to 0.1. During inference, responses are generated using nucleus sampling with top-$p = 0.8$, a temperature of $\tau = 0.3$, and a maximum sequence length of 1024 tokens.

D Ablation Study

Due to space constraints in the main text, we present the comprehensive ablation study results on the MSC and GC datasets in this section. Table 8 details the performance of our proposed ThinkFlow and its variants on these two additional datasets. Consistent with the observations on the CC dataset, removing any key architectural component (PLMS, GLC, or CAHA) or training phase (Phase1 or Phase2) leads to a performance drop across most metrics. This further demonstrates the robustness of our design choices and the generalizability of our model's capabilities across different evaluation domains.

Algorithm 1: ThinkFlow: Self-Supervised Test-Time Evolution

Input: Long-term dialogue stream $\mathcal{D} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_N\}$, Teacher LLM $\mathcal{M}_T$, Student LLM $\mathcal{M}_S$ (with LoRA), Memory modules (PLMS, GLC, CAHA)
Output: Optimized Student LLM $\mathcal{M}_S$ and Memory modules

Initialize latent memory skills $S_0 \leftarrow \emptyset$;
`/* Phase 1: Teacher-Guided Latent Alignment */`
for *epoch* $e = 1$ **to** E **do**
  $S_{t-1} \leftarrow S_0$;
  for *each turn* (x_t, a_t) *in cold-start session* $\mathcal{S}_1$ **do**
    $P_T \leftarrow \mathcal{M}_T(y \mid H_{<t}, x_t)$; `// Target dist. from explicit history`
    $S_{aligned} \leftarrow \text{CAHA}(S_{t-1}, x_t)$; `// Dynamic latent translation`
    $P_S, E_t \leftarrow \mathcal{M}_S(y \mid S_{aligned}, x_t)$; `// Student prediction`
    $\hat{S}_{new}, \mathcal{L}_{KL} \leftarrow \text{PLMS}(E_t)$; `// Compress context`
    $S_t \leftarrow \text{GLC}(\hat{S}_{new}, S_{t-1})$; `// Consolidate memory`
    $\mathcal{L}_{Phase1} \leftarrow \mathcal{D}_{KL}(P_T \parallel P_S) + \alpha \mathcal{L}_{KL}$;
    $\nabla \mathcal{L}_{Phase1} \rightarrow$ Update $\mathcal{M}_S$, PLMS, GLC, CAHA;
  end
end
`/* Phase 2: Next-User-Utterance Prediction (Online Evolution) */`
for *each session* $\mathcal{S}_i \in \{\mathcal{S}_2, \ldots, \mathcal{S}_N\}$ **do**
  for *each turn* (x_t, a_t) *and next user query* x_{t+1} *in* $\mathcal{S}_i$ **do**
    $S_{aligned} \leftarrow \text{CAHA}(S_{t-1}, x_t)$;
    $E_t \leftarrow \mathcal{M}_S(S_{aligned}, x_t, a_t)$;
    $\hat{S}_{new}, \mathcal{L}_{KL} \leftarrow \text{PLMS}(E_t)$;
    $S_t \leftarrow \text{GLC}(\hat{S}_{new}, S_{t-1})$;
    $P_{pred} \leftarrow \mathcal{M}_S(x_{t+1} \mid S_t, x_t, a_t)$; `// Implicit prediction feedback`
    $\mathcal{L}_{CE} \leftarrow \text{CrossEntropy}(P_{pred}, x_{t+1})$;
    $\mathcal{L}_{Conf} \leftarrow -\sum P_{pred} \log P_{pred}$; `// Min-Entropy regularization`
    $\mathcal{L}_{Phase2} \leftarrow \mathcal{L}_{CE} + \beta \mathcal{L}_{Conf} + \alpha \mathcal{L}_{KL}$;
    $\nabla \mathcal{L}_{Phase2} \rightarrow$ Update $\mathcal{M}_S$, PLMS, GLC, CAHA;
  end
end

E Prompts

Unlike traditional explicit memory pipelines that suffer from the text bottleneck by appending retrieved text summaries, ThinkFlow maintains a minimal textual context (Figure 6). Instead, historical user states are dynamically injected as continuous probabilistic latent memory skills directly into the embedding space.

F Parameter Analysis

Scaling latent capacity is increasingly vital for ultra-long interactions to prevent feature blending. Figure 7 analyzes the scaling effect of latent memory skills (K) across 32K, 128K, and 1M context lengths. While

Category	Hyper-parameter	Value
LoRA	Target Modules	`q_proj`, `k_proj`, `v_proj`, `o_proj`
	Rank (r)	8
	Alpha (α)	16
Memory Architecture	PLMS Attention Heads	8
	GLC MLP Intermediate Size	$D/2$
	CAHA Bottleneck Rank	8
	Number of Skills (K)	10
Training	Optimizer	AdamW
	Learning Rate	2×10^{-5}
	Max Gradient Norm	1.0
	KL Divergence Weight (α)	0.1
	Min Entropy Weight (β)	0.1
	Cold-Start Epochs	3
Inference	Decoding Strategy	Nucleus Sampling
	Top-p	0.8
	Temperature (τ)	0.3
	Max Sequence Length	1024

Table 7: Detailed hyper-parameter configurations for ThinkFlow.

Methods	MSC				GC			
	B-4	**R-L**	**Bert**	**Mauve**	**B-4**	**R-L**	**Bert**	**Mauve**
ThinkFlow (Ours)	**1.12**	**14.33**	**46.72**	**66.48**	**1.10**	**14.21**	44.82	65.26
w/o PLMS	0.96	13.33	45.06	65.09	0.93	9.32	37.88	45.19
w/o GLC	1.03	13.70	45.17	65.83	1.03	13.70	**45.17**	**65.83**
w/o CAHA	1.10	14.21	44.82	65.26	1.04	10.25	39.89	52.23
w/o Phase1	1.07	13.98	44.71	64.22	1.07	13.98	44.71	64.22
w/o Phase2	1.09	14.03	44.86	64.35	1.09	14.03	44.86	64.35

Table 8: Ablation study of our method (Qwen3-8B) on MSC and GC datasets, where w/o means without.

Prompt for ThinkFlow

```
"""
<|im_start|>system
Generate the most plausible next response like a human based on the current conversation. Do not put too
much information in the next response.<|im_end|>

[Latent Memory Skills S Injected Here as Continuous Embeddings]

<|im_start|>user
{user_utterance}<|im_end|>

<|im_start|>assistant
"""
```

Figure 6: Prompt structure for generating personalized agent responses in ThinkFlow.

shorter contexts (32K) maintain relatively high accuracy even with fewer skills, the ultra-long 1M context scenario exhibits a sharp and robust upward trend as K increases from 1 to 10. This reveals a critical insight: as the dialogue history grows exponentially, allocating a larger set of disentangled latent vectors becomes essential to independently capture expanding user nuances without semantic interference.

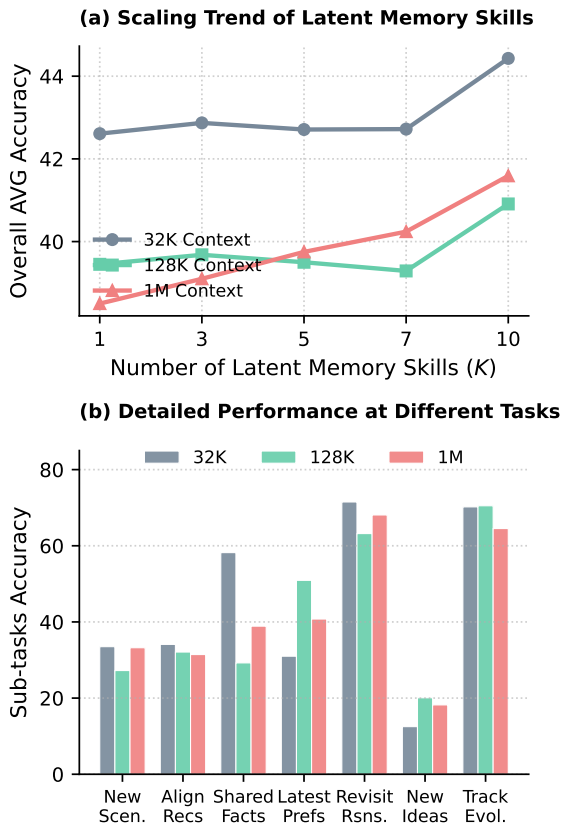

Figure 7: Latent memory scaling on PersonaMem.

G Algorithm Walkthrough

Algorithm 1 outlines the two-phase training paradigm of ThinkFlow. **Phase 1 (Lines 2-12)** addresses the cold-start problem by distilling knowledge from a text-based Teacher LLM. The teacher accesses the full explicit context to generate a target distribution. In contrast, the Student LLM relies solely on the dynamically translated latent memory skills (via CAHA). The student updates its latent state through the PLMS and GLC modules and aligns with the teacher by minimizing the KL divergence. **Phase 2 (Lines 13-25)** enables continuous, label-free evolution. Operating purely in the latent space, the agent uses implicit predictive feedback to anticipate the user's next utterance (x_{t+1}) based on the consolidated memory. The framework continuously optimizes itself through a cross-entropy loss against the actual user reaction, regularized by a minimum entropy confidence prior.